Sensitivity Analysis for Predicting Sub-Micron Aerosol Concentrations Based on Meteorological Parameters

Zaidan, Martha A.

2020-05

Zaidan , M A , Surakhi , O , Fung , P L & Hussein , T 2020 , ' Sensitivity Analysis for Predicting Sub-Micron Aerosol Concentrations Based on Meteorological Parameters ' , Sensors , vol. 20 , no. 10 , 2876 . https://doi.org/10.3390/s20102876

http://hdl.handle.net/10138/323391
https://doi.org/10.3390/s20102876

sensors

MDPI

Article

Sensitivity Analysis for Predicting Sub-Micron Aerosol Concentrations Based on Meteorological Parameters

Martha A. Zaidan [1,*], Ola Surakhi [2], Pak Lun Fung [1] and Tareq Hussein [1,3,*]

1 Institute for Atmospheric and Earth System Research (INAR)/Physics, University of Helsinki, FI-00560 Helsinki, Finland; pak.fung@helsinki.fi
2 Department of Computer Science, The University of Jordan, Amman 11942, Jordan; ola.surakhi@gmail.com
3 Department of Physics, The University of Jordan, Amman 11942, Jordan
* Correspondence: martha.zaidan@helsinki.fi (M.A.Z.); tareq.hussein@helsinki.fi or t.hussein@ju.edu.jo (T.H.)

Received: 24 April 2020; Accepted: 15 May 2020; Published: 19 May 2020

Abstract: Sub-micron aerosols are a vital air pollutant to be measured because they pose health effects. These particles are quantified as particle number concentration (PN). However, PN measurements are not always available in air quality measurement stations, leading to data scarcity. In order to compensate this, PN modeling needs to be developed. This paper presents a PN modeling framework using sensitivity analysis tested on a one year aerosol measurement campaign conducted in Amman, Jordan. The method prepares a set of different combinations of all measured meteorological parameters to be descriptors of PN concentration. In this case, we resort to artificial neural networks in the forms of a feed-forward neural network (FFNN) and a time-delay neural network (TDNN) as modeling tools, and then, we attempt to find the best descriptors using all these combinations as model inputs. The best modeling tools are FFNN for daily averaged data (with $R^2 = 0.77$) and TDNN for hourly averaged data (with $R^2 = 0.66$) where the best combinations of meteorological parameters are found to be temperature, relative humidity, pressure, and wind speed. As the models follow the patterns of diurnal cycles well, the results are considered to be satisfactory. When PN measurements are not directly available or there are massive missing PN concentration data, PN models can be used to estimate PN concentration using available measured meteorological parameters.

Keywords: particle number concentration; modeling; sensitivity analysis; artificial neural networks; feed-forward neural network; time-delay neural network

1. Introduction

1.1. Motivation

Approximately seven million people die every year due to adverse health-related air pollution issues, in which 4.2 million deaths are attributed to exposure to poor outdoor air quality. Approximately 91% of the world's population lives in areas where air pollution exceeds guideline limits established by the World Health Organization (WHO) [1]. The most critical air pollutants from a health perspective include airborne particulate matter (PM) and the gaseous pollutants, such as ozone (O_3), nitrogen dioxide (NO_2), volatile organic compounds (e.g., benzene), carbon monoxide (CO), and sulfur dioxide (SO_2) [2,3]. In particular, particles less than 2.5 micrometers in diameter ($PM_{2.5}$) are able to penetrate deeply into human lungs, irritate and corrode the alveolar wall, and consequently impair lung function. Although the diameter of $PM_{2.5}$ is very small, it has a large surface area, and then may be capable of carrying various toxic substances, passing through the filtration of nose hair, reaching the end of the respiratory tract with airflow, and accumulating there by

diffusion, damaging other parts of the body through air exchange in the lungs [4]. Atmospheric PM also plays a role in ecosystems and Earth's climate, leading to extensive research on the subject [5].

A critically important class of atmospheric PM is called ultra-fine particles (UFPs). These particles are smaller than 100 nm in size (i.e., sub-micron aerosols). Scientific attention has recently moved toward UFPs because these particles have very high surface area to mass ratios. Consequently, they can easily enter the human respiratory system and deposit preferentially in the deepest areas of the lungs, such as the tracheobronchial and alveolar regions, carrying toxic compounds [6]. Emissions associated with traffic, industrial activities, and domestic heating contribute to a large fraction of UFPs [7]. Particle number (PN) concentrations are more informative in describing the abundance of UFPs because these particles tend to dominate atmospheric PM number size distributions and contribute little to PM mass concentrations that are presently used as air quality indicators (e.g., $PM_{2.5}$ and PM_{10}) [8]. Unfortunately, there are much fewer data available on PN compared with PM [9,10] due to the unavailability of instruments for measuring UFP in many air quality monitoring stations [6]. Therefore, we propose in this paper a modeling framework to be an alternative method in estimating PN concentration using other available measurements. In this way, PN concentration can be monitored in cities where the measurements are not available, and the air quality database can be updated for further analysis.

1.2. Data-Driven Air Pollutant Modeling

Modeling air pollutants can generally be categorized into three main approaches, including: physics- and expert-based and data-driven approaches [11]. First, physics-based approaches use models that describe underlying physical processes related to air pollutants directly [12]. This modeling approach is typically accurate and reliable, but physics and chemistry knowledge is required, especially related to a particular air pollutant to be modeled. In some cases, they can be computationally demanding and may also be sensitive to the scale and quality of the parameters involved [13]. Examples are the urban airshed model (UAM) [14] and the community multiscale air quality (CMAQ) model [15]. Second, expert-based approaches, such as the expert elicitation process [16], elicit knowledge from experts/specialists for modeling and analysis [17]. The involvement of experts may be helpful to explain data anomalies or pattern outliers due to untypical air pollution phenomena, such as forest fires, sudden traffic changes, etc. However, it is often difficult to find agreement among experts about the use of expert systems and how the uncertainties of different variables can be adequately accounted for [16]. Finally, the data-driven approach uses historical datasets to identify relationships between measured variables and then builds models based on the trends in the data. This approach does not typically require deep knowledge in air pollutant dynamics, chemistry composition, and other explanatory variables. Due to these reasons, more practitioners have recently utilized data-driven approaches, such as neural networks, as alternatives to physics- and expert-based methods, to model air pollutant concentrations [18]. This work resorts to a data-driven approach in the form of artificial neural networks (ANN) to model and estimate PN concentrations. In particular, sensitivity analysis is carried out to find the best combination of measured variables for estimating PN concentrations.

Data-driven-based modeling has been carried out for estimating different air pollutant concentrations, including nitrogen dioxide (NO_2) [19], sulfur dioxide (SO_2) [20,21], ozone (O_3) [22,23], black carbon [11,24], particulate matter smaller than 10 μm (PM_{10}) [21,25], and particulate matter smaller than 2.5 μm ($PM_{2.5}$) [26–28]. However, there is a very limited number of studies focusing on estimating PN concentration. The estimations of PN concentration using data-driven methods were focused on European cities, described in [29,30]. For the first time, this work proposes a data-driven framework for estimating PN concentration in the Middle East and North Africa (MENA) region. Furthermore, the modeling framework evaluates the performance by applying different combination of measured variables, which is known as sensitivity analysis. The best combination of measured variables leads to reliable PN models and then allows filling in the missing data in the air

quality database and estimating the PN concentration without relying on expensive measurements. The capability to estimate PN concentration on a daily and hourly basis allows a decision maker, such as a government agency, to mitigate the impact caused from these sub-micron aerosols.

2. Materials

This section describes the materials used in this study. We explain the experimental setup, and then, we describe how the data were pre-processed. We also discuss the environmental conditions during the measurement period.

2.1. Database

In this study, we used a database obtained from a measurement campaign in Amman, the capital city of Jordan, from 1 August 2016 until 31 July 2017. The city is considered as an area with Middle Eastern urban conditions within the MENA region. This region serves as a compilation of different aerosol particle sources including natural dust, anthropogenic pollution (e.g., generated from the petrochemical industry and urbanization), as well as new particle formation [31].

The database includes sub-micron particle number concentration (PN) and meteorological conditions. The aerosol measurement was performed at the aerosol laboratory, which is located on the third floor of the Department of Physics, University of Jordan. The campus is located in an urban background in the north part of Amman, Jordan. In particular, the campaign measured the particle number size distribution using a scanning mobility particle sizer (NanoScan SMPS 3910, TSI, MN, USA). The time resolution used in the SMPS was 1 min. The meteorological measurement was performed with a weather station (WH-1080, Clas Ohlson: Art.no.36-3242, Helsinki, Finland) with a 5 min time resolution. The meteorological data were comprised of ambient temperature (T), absolute pressure (P), relative humidity (RH), wind speed (WS), and wind direction (WD). The details of the aerosol measurement campaign and the meteorological measurements were described in [31,32].

2.2. Data Handling

The particle number concentrations (PN), in cm^{-3}, were calculated by integrating the measured particle number size distribution over the specified particle diameter range, given by:

$$\mathrm{PN_{sub}} = \int_{10\,\mathrm{nm}}^{450\,\mathrm{nm}} n_N^0 \, dlog_{10}(Dp) \tag{1}$$

where $n_N^0 = dN/dlog_{10}(Dp)$ is the measured particle number size distribution and Dp is the particle diameter. Since air quality data are typically reported hourly or daily, the processed aerosol data (PN_{sub} concentration) and meteorological measurements were averaged hourly and daily. Having the data for a year at an hourly resolution allowed the modeling to capture the diurnal cycle and seasonal variability.

2.3. Environmental Conditions

Figure 1a shows time-series data of PN during the campaign. The red curve represents the daily average, whereas the blue curve indicates the hourly measurement. It can be seen that the PN concentration ranged between 10^3 cm^{-3} and 10^5 cm^{-3}, with median values of 1400 cm^{-3} and 1500 cm^{-3}, for daily and hourly data, respectively. In addition, Figure 2 presents PN histograms for daily (left subplot) and hourly (right subplot). It can be seen that both histograms peaked at the bin edge at about 1330 cm^{-3}. Understanding the ranges and the median values of PN_{sub} concentrations allowed us to examine later if the modeling metrics were adequate.

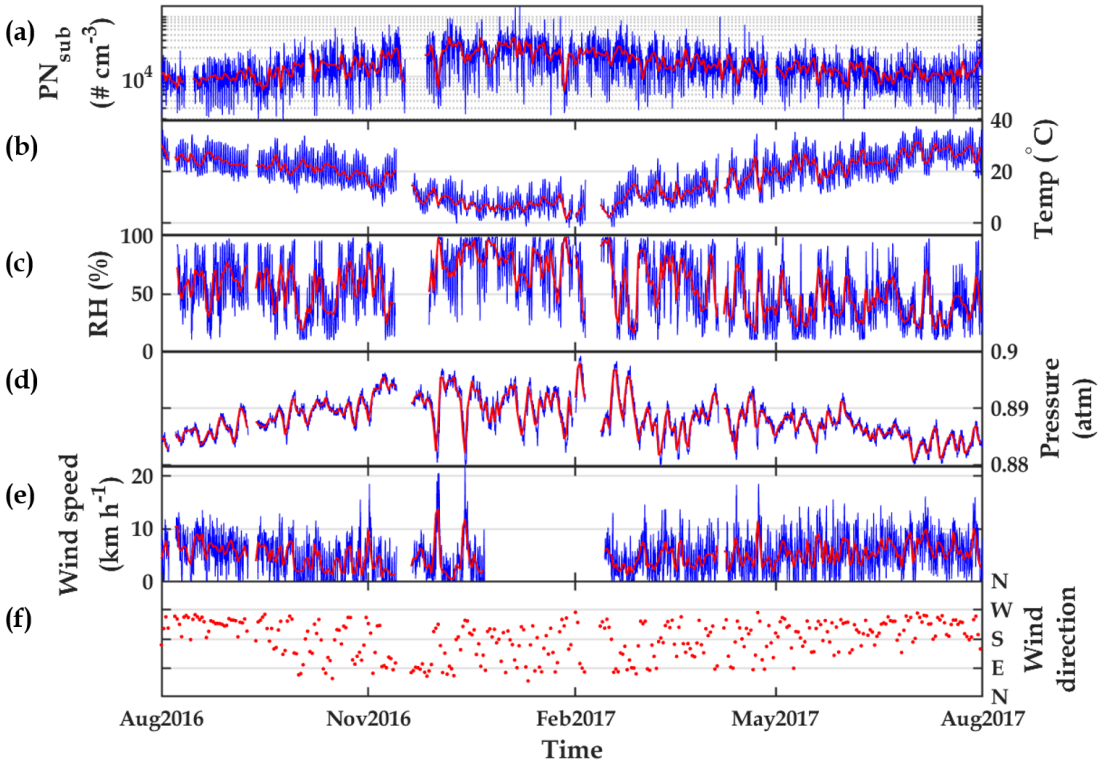


Figure 1. PNsub, shown in subfigure (**a**), and meteorological conditions, shown in subfigures (**b**–**f**), during the experiments. Red and blue colors are daily and hourly averaged.

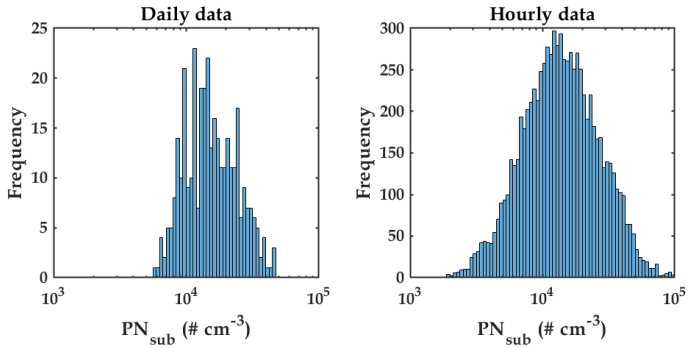


Figure 2. PN number concentration histograms during the experiments for daily averaged data (**left**) and hourly averaged data (**right**).

Figure 1b–f present the meteorological conditions during the experiment. The red and blue curves indicate daily and hourly averaged data. Figure 1b shows the temperature (T), with a minimum peak of about 0 °C during winter and a maximum of about 40 °C during summer with the hourly median value of 19.9 °C. Figure 1c indicates the relative humidity (RH), which varied between 10% and 100% with the hourly median value of 52.3%. Figure 1d is the pressure (P), which ranged between 0.88 atm and 0.9 atm, and its median value was 0.888 atm. Figure 1e indicates wind speed (WS), ranging between 0 km/h and 20 km/h with a median value of 5 km/h. It can be seen that there were about 2 months of missing data in this variable. Finally, Figure 1f represents wind direction (WD), where only daily averaged data are shown for better visualization. Wind blows mainly from the south and west (180°–270°) from June to September. The wind direction varies in other months, ranging from 45° to 270°.

3. Methods

This section describes the methodology for estimating PN concentration used in this study. Figure 3 shows a block diagram illustrating this methodology. First, a database was formed using processed data from aerosol and meteorological measurements as described in Section 2.2. In the second step, the data underwent pre-processing procedures through data cleaning and data normalization, which will be explained in Section 3.1. The next step was a part of the sensitivity analysis block, consisting of several sub-steps.

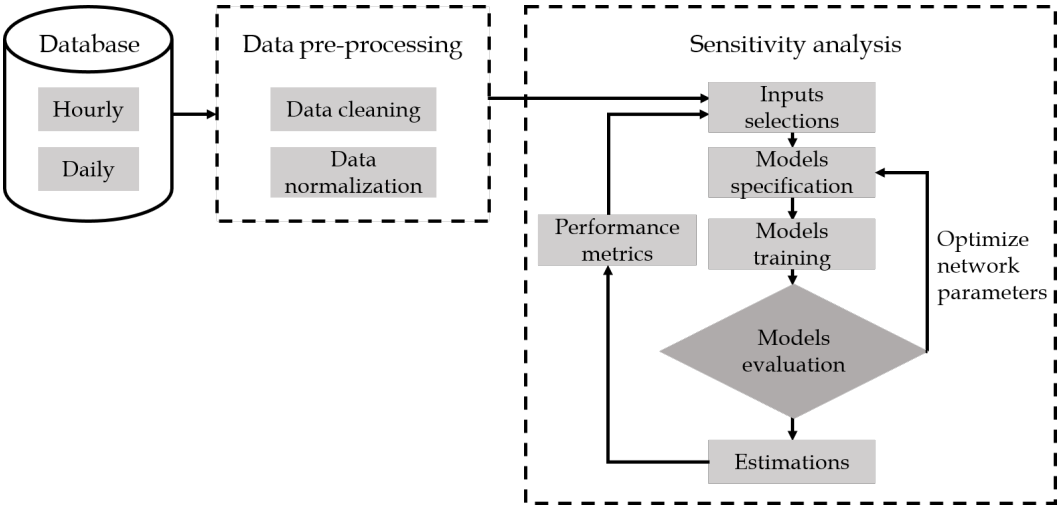


Figure 3. The block diagram of the proposed sensitivity analysis for estimating PN concentration.

In this work, sensitivity analysis could be defined as a methodology to find the best combination of measured variables for modeling PN concentration. Sensitivity analysis is more effective than performing bivariate correlation analysis, including linear correlation, such as Pearson [33] and Spearman [34], and non-linear correlation analysis, such as mutual information [35]. Bivariate correlation analysis is beneficial when two variables are investigated in terms of their relationship, but when there are more than two variables interacting in multivariate directions, those methods may no longer be effective. The first sub-step in sensitivity analysis is input selection. This sub-step prepares a set of different measured variable combinations. Every single combination is then fed as inputs for a chosen data-driven model. The next sub-step is to specify the chosen model structure and other model properties. Then, the model parameters can be optimized in the model training sub-step. Once the model parameters have been optimized, the model is then evaluated using selected metrics; if the performance is not satisfactory, the model structure needs to be re-specified. These steps are done iteratively until we achieve satisfactory performance defined by a modeler. Once the model has met satisfactory condition, it estimates PN concentration using test data. Finally, performance metrics can be evaluated, and the next sub-step is to take other input combinations. These sub-steps can be done in sequence or in parallel depending on available computing resources.

3.1. Data Pre-Processing

The aerosol measurement data obtained from 1 August 2016 to 31 July 2017 were processed to give the PN concentration. The meteorological measurements (T, RH, P, WS, WD) were collected for the same period. Both data were merged, and the data were averaged daily and hourly, resulting in 365 and 8760 observations, respectively. The data pre-processing began by removing the missing data. The missing data constituted 6.7% of the total data points due to technical faults or instrument maintenance. Since the data were obtained from different measured variables with various physical units and magnitude, it was crucial to normalize the data. The scaling factor depended on the chosen

model, which contained activation functions ranging between these values. In this case, the data were scaled between 0 and 1 to transform them into the range of the activation function.

3.2. Modeling

The first sub-step in sensitivity analysis is to prepare a set of measured variable combinations, as shown in Figure 4. Then, the inputs for a chosen model are selected based on this combination, which can be done in sequence or in parallel.

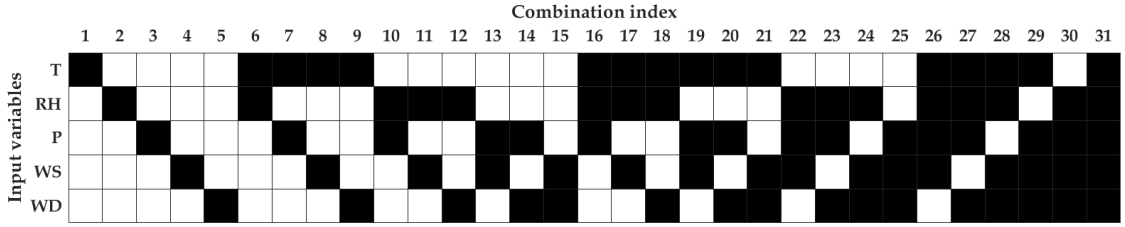


Figure 4. Sensitivity analysis uses different combinations of meteorological variables as inputs for PN modeling.

The second sub-step is to prepare a PN model. In this case, we used an artificial neural network (ANN) to model PN concentration. Neural networks (NNs) provide a robust approach for approximating real-valued (regression) and discrete-valued (classification) target functions because they can mimic the non-linearity of the functions and their optimization methods are well developed [36]. These models have been a popular choice among machine learning methods for approximating complex functions [37] and have been utilized in a large number of applications [38], including air pollution [18] and climate [39]. In this case, we resorted to two types of NNs, which were a feed-forward neural network (FFNN) and a time-delay neural network (TDNN). FFNN is a fully-connected network with two layers (input and hidden layers). FFNN has been the mostly popular choice of NNs due to its fast operation, ease of implementation, and smaller training set requirements [40]. TDNN structure is the same as FFNN, but the feed-forward network has a tapped delay line at the input. TDNN is part of a general class of dynamic networks, where the dynamics appear only at the input layer of a static multi-layer feed-forward network. This type of network is suited well for dealing with time-series data [41].

Both FFNN and TDNN estimate PN concentration, $\hat{\mathbf{y}}$, through the function of meteorological variables, $f(\mathbf{x}, \mathbf{w})$, by optimizing the weights, $\mathbf{w}$, of NN. Figure 5 displays a schematic representation of a neural network with one hidden layer. The jth neuron in the Lth layer calculates the output z_j^L as:

$$z_j^L = \sigma\left(\sum_i w_{ji}^L x_i + b_j^L\right) \tag{2}$$

where the notation w_{ji}^L represents the weight of connection between the computing neuron and its ith input in the preceding layer and b_j^L is a bias parameter. In the case of TDNN, a tapped delay line is introduced at the input layer, where the input data are buffered for several time steps and then fed to the input layer. The introduction of time delays (T) allows each neurons to have access to n input values, corresponding to different input array instantaneous responses $x(t - nT), \ldots, x(t)$. The symbol $\sigma(.)$ is the activation function in the hidden layer. In this case, we used the rectified linear unit (ReLU) activation function in the first layers (i.e., input and hidden layers), whereas the linear activation function was used in the output layer. Once a training dataset, $\{\mathbf{x}, \mathbf{y}\}$, with reference

inputs, **x**, and their corresponding outputs, **y**, was provided, optimized weights, **w**, could be found by minimizing the cost function:

$$E = \sum_{n} \left(f(\mathbf{x}_n, \mathbf{w}) - \mathbf{y}_n \right)^2 \tag{3}$$

where $f(\mathbf{x}_n, \mathbf{w})$ is the output of the NN from the training inputs $\mathbf{x}_n$. The optimization was done through stochastic gradient descent. This sub-step is called model training. In the next sub-step, the model was also evaluated to observe if the model specification was satisfactory. This step was done iteratively through k-fold cross-validation, which is a resampling technique designed to partition dataset into k (k-fold) subsets of data where one sample of them is held out while the model is trained with the remaining samples and then tested on the hold-outs. Iteration is also carried out to find the best model configurations, by adjusting the number of neurons for the input and hidden layers, weight initiation, the number of training cycles (epochs), and the learning rate.

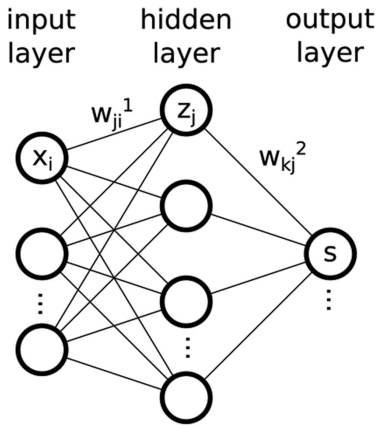


Figure 5. Schematic representation of a neural network with one hidden layer.

Once the model met a satisfactory performance defined by a modeler, the testing data could be fed into the trained network to estimate PN concentration. The results were then evaluated through several performance metrics, which will be explained in the following sub-section.

3.3. Performance Metrics

In order to evaluate the best PN model through sensitivity analysis, we resorted to three metrics, as shown in Table 1. The symbols y, $\bar{y}$, and $\hat{y}$ represent the real measurement value, the mean of the measurement data points, and the estimated model value, respectively. The point number and the total estimated values from the models are indicated by the notations i and n, respectively. The coefficient of determination (R^2) provides a measure of how well the observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model. The mean absolute error (MAE) gives a simple interpretation as the average absolute difference between the predicted model values ($\hat{\mathbf{y}}$) and the real measurement data points (**y**). Root mean squared error (RMSE) represents the standard deviation of the estimated errors (i.e., error residuals).

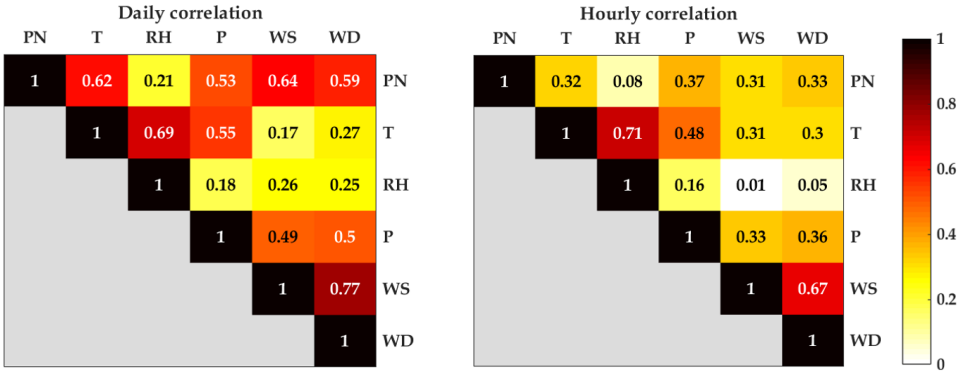
Table 1. The performance metrics used in the sensitivity analysis for the PN models' evaluation.

Performance Metrics	Formulation
Coefficient of Determination	$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$
Mean Absolute Error	$MAE = \frac{\sum_{i=1}^{n}\lvert\hat{y}_i - y_i\rvert}{n}$
Root Mean Squared Error	$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$

4. Results

4.1. Data Analysis

Figure 6 shows two matrix plots indicating the level of absolute Pearson correlation coefficients (PCC) between measured variables for daily and hourly data. The color closest to light yellow indicates a weak correlation, whereas the color closest to black indicates a strong correlation. It can be seen that the daily RH had a modest correlation with PN (PCC was about 0.21, with a *p*-value equal to zero), whereas the remaining daily measured variables had PCC values greater than 0.5, which indicated good correlations with PN. The hourly PCC values seemed to be reduced when compared to daily average data. The hourly RH showed a very weak correlation with PN (with PCC lower than 0.1, with a *p*-value equal to 0.54). Other meteorological variables, such as T, P, WS, and WD, still demonstrated satisfactory correlation with PN, ranging between 0.31 and 0.37 (with a *p*-value equal to zero).

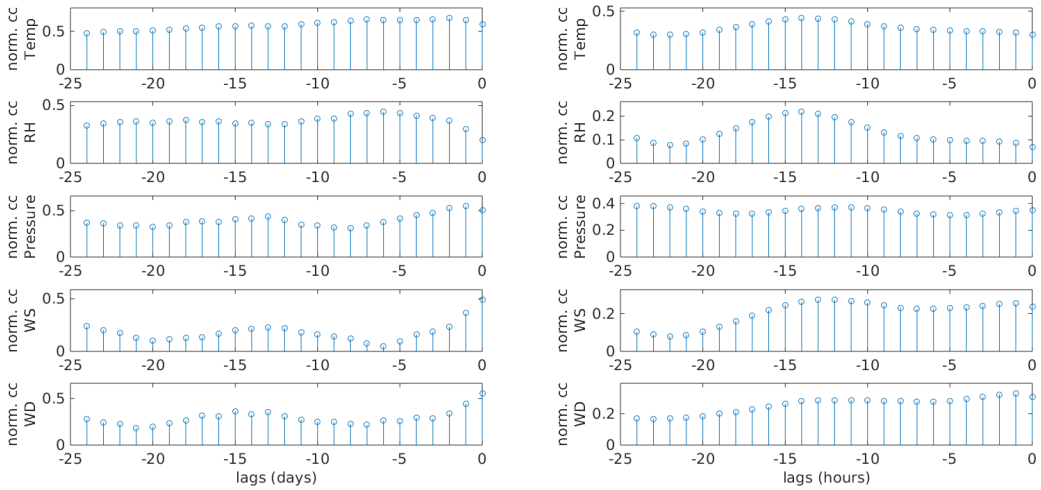


(**a**) (**b**)

Figure 6. Matrix plots: absolute Pearson correlation coefficients between measured variables for daily and hourly averaged data. (**a**) Daily. (**b**) Hourly.

Figure 7 shows the cross-correlation between PN and meteorological variables for daily (Figure 7a) and hourly (Figure 7b) averaged data. The x-axis shows different time lags, and the y-axis represents normalized correlation coefficients (abbreviated as norm. cc in the Figure). Both sub-figures demonstrate clearly that previous meteorological variables influenced the current PN concentration. Therefore, the use of time-delayed meteorological measurements may be beneficial in improving PN modeling accuracy based on the hourly data.

In general, the use of a large number of inputs typically increases the model complexity, leading to limited model performance. On the other hand, limiting the number of inputs also allows a model to be used without depending on many other measurements in practice. Therefore, it is vital to consider these effects when determining the number of inputs involved in modeling [23]. Since the matrix plotted only indicated bivariate correlation analysis, i.e., the correlation between two variables, sensitivity analysis was a useful method to investigate the multivariate measured variables influencing PN concentration. Sensitivity analysis was performed by training and testing PN modeling on all possible measured variable combinations, then the best combination of measured variables explaining PN concentration could be used as a final PN model.



(**a**) (**b**)

Figure 7. Cross-correlation between PN data and meteorological parameters for daily and hourly averaged data. Different time lags are shown on the x-axis, whereas the y-axis represents normalized correlation coefficients (norm. cc). (**a**) Daily. (**b**) Hourly.

4.2. Sensitivity Analysis

We resorted to two different types of ANN, called FFNN and TDNN, with the model specifications mentioned in Section 3. These models were then tested on all combination indexes of measured meteorological variables to perform sensitivity analysis, as illustrated in Figure 4. The models were trained and tested twice using daily and hourly average data. In this way, several best combinations of meteorological variables could be evaluated, and then, the best combination would be selected to be used as the inputs of PN models. The number of combinations of the variables used was 5 (if one variable was used), 10 (if two variables were used), 10 (if three variables were used), 5 (if four were variables used), and 1 (if five variables were used), with the total combinations being 31.

The performance metrics of modeling using these input variable combinations were tested according to R^2, MAE, and RMSE. Figures 8 and 9 present the performance metrics of PN modeling for daily and hourly averaged data, respectively. The blue bars are FFNN, whereas the red bars are TDNN. The low values of MAE and RMSE indicated that the models' performance was better than the high values of these metrics. On the other hand, the high R^2 values indicated that the models' performance was better than the lower values. Since there were three metrics involved, the first priority was given to R^2, then MAE and RMSE.

It can be seen that for both models, i.e., FFNN and TDNN, applied on both types of data averaging, having much fewer inputs did not provide adequate model performance because the inputs used were not informative enough to describe the PN concentration. As a general rule, having more variables for the inputs increases modeling accuracy.

Through the evaluation of the R^2 values, Figure 8 (daily data averaging) shows that both models performed well when the models used at least four measured variable combinations. From these, the best R^2 values were found at the combination indexes of 26 and 31. R^2 values for FFNN were found to be the same for both models, that is equal to 0.78. However, the R^2 value for TDNN for the combination index 26 (R^2 = 0.77) was better than Number 31 (R^2 = 0.71). Therefore, we decided to use Combination Index 26 for daily PN modeling with T, RH, P, and WS as input variables. In particular, FFNN seemed to be better than TDNN by observing R^2 and RMSE values. Although MAE showed otherwise, we decided to resort to FFNN as the PN model because of the simplicity of model's specification, development, and usage. Overall, when there were more than four inputs involved, FFNN also provided better performance than TDNN.

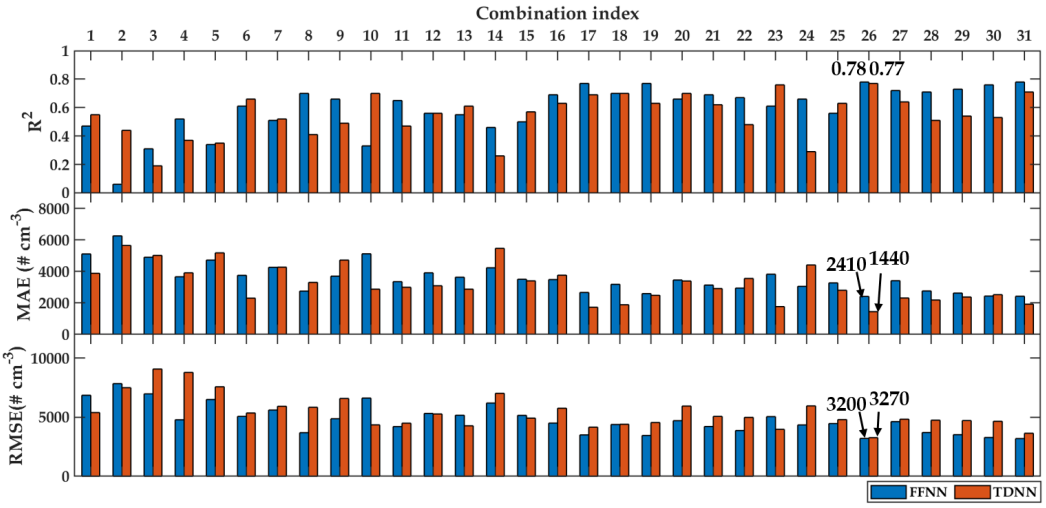


Figure 8. Performance metrics of daily modeling using FFNN (blue) and TDNN (red). The top, middle, and bottom sub-figures are R^2, MAE, and RMSE, respectively.

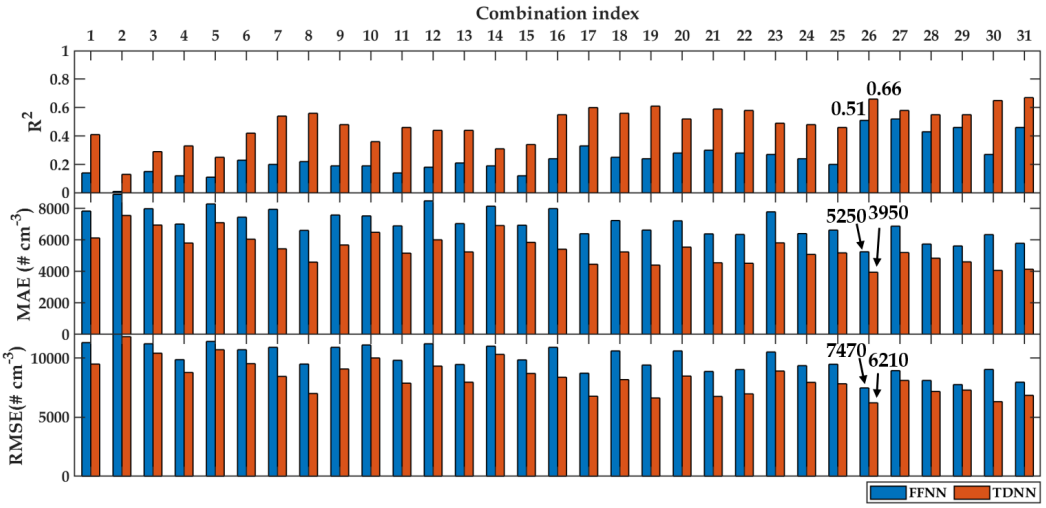
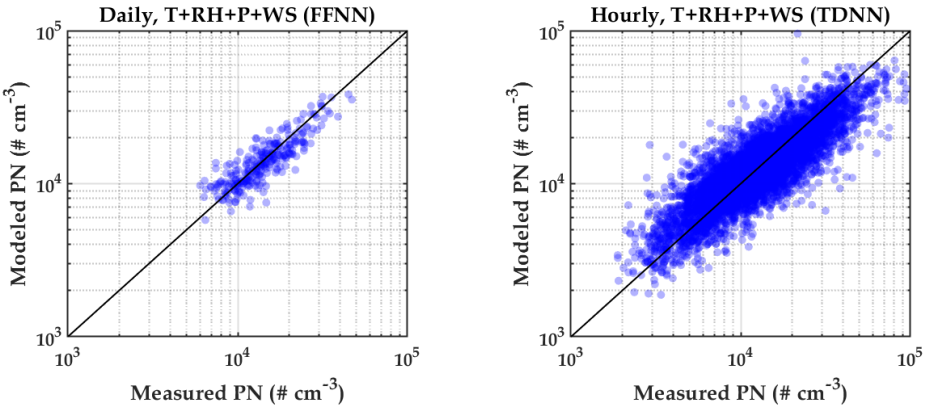


Figure 9. Performance metrics of hourly modeling using FFNN (blue) and TDNN (red). The top, middle, and bottom sub-figures are R^2, MAE, and RMSE, respectively.

On the other hand, Figure 9 demonstrates clearly that TDNN was better than FFNN in all performance metrics across all input combinations. As in the case of daily data averaging, the best two candidates were found in the combination indexes 26 and 31. However, the R^2 value of the combination index 31 (R^2 = 0.67) was slightly better than 26 (R^2 = 0.66). In this case, we decided to use the combination index 26 due to several reasons. First, the combination index 26 was found to be in agreement with the modeling using daily averaged data. Second, the combination index 26 had the best performance in terms of the MAE and RMSE metrics. Third, the best performing model 26 excluded the WD variable. WD was a circular variable, and in this study, we showed it in the scale of 0° to 360°, which created discontinuity at the north. To tackle this, a trigonometric function had to be applied to resolve WD into two perpendicular directions before the data analysis. Finally, it was better to have a model that used fewer input variables if the performance was similar to a model with additional inputs. Therefore, in practice, the model with fewer inputs relied on fewer measurements (i.e., instruments). In summary, both models (using daily and hourly averaged data) used the combination index 26 with measured variables of T, RH, P, and WS. For now on, we present the results of PN models using the combination index 26.

Figure 10 shows the scatter plots between the reference measured PN and the modeled PN using variables T, RH, P, and WS, for daily (Figure 10a) and hourly (Figure 10b) averaged data. It can be

seen that most estimated data points of PN concentration for both averaged data followed the 1:1 line. Figure 11 shows the histograms of residual error between the PN measurement and the PN model for daily (left) and hourly (right). The figure shows that the majority of residual data points lied around zero residual error. This indicated that most of the estimated PN data points were precise.



(**a**) (**b**)

Figure 10. Scatter plot between PN measurement and PN estimation using four measured meteorological variables (T, RH, P, and WS). (**a**) Daily. (**b**) Hourly.

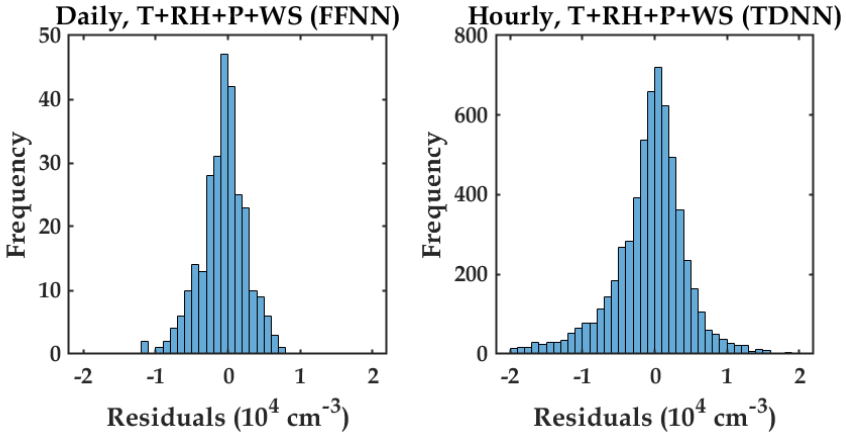


Figure 11. Histograms of residual error between the reference instrument and PN estimation using four measured variables (T, RH, P, and WS).

Figure 12 presents the median of diurnal cycles calculated on all days in a week for measured PN (blue) and modeled PN (red). Furthermore, Figure 13 shows the median of diurnal cycles calculated on workdays and weekends for measured PN (blue) and modeled PN (red). Both figures show that the PN model followed the patterns of diurnal cycles in these two scenarios well. The results emphasized that PN modeling, using TDNN with input variables of T, RH, P, and WS, was reliable. Even though the variable RH presented a weak correlation to PN as shown in Figure 6a,b, through sensitivity analysis, it was found that it became a good feature when combined together with other variables. This indicated that using bivariate correlation analysis may not be optimal to find the best combination of variables to estimate PN concentration. Therefore, sensitivity analysis was a suitable method in this case.

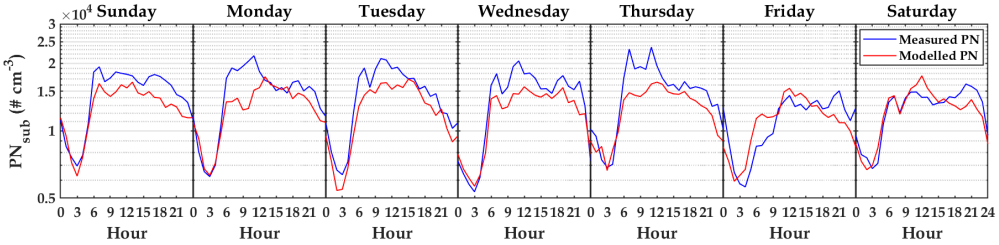


Figure 12. The median of diurnal cycles calculated on different days for measured PN and modeled PN (No. 26).

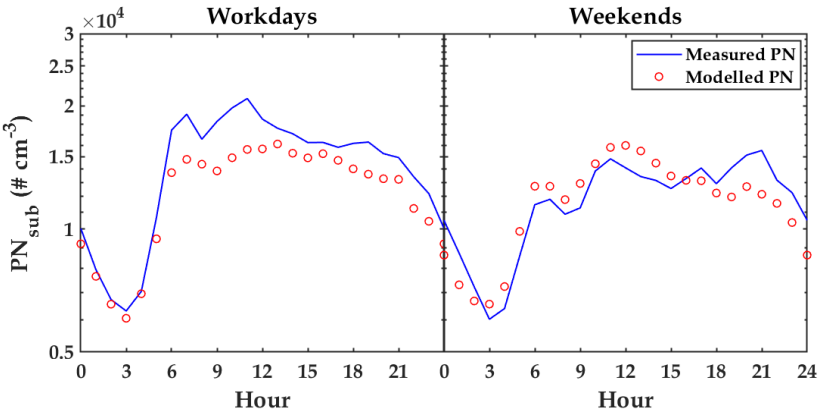


Figure 13. The median of diurnal cycles calculated on workdays and weekends for measured PN and modeled PN (No. 26).

5. Conclusions

Due to the adverse health effects on the human respiratory system, PN concentration is a vital air pollutant to be measured or estimated if the measurement is not available or there are massive missing data. This paper presented applying sensitivity analysis in the modeling framework of PN estimation. A feed-forward neural network (FFNN) and a time-delay neural network (TDNN) were chosen as the PN modeling tools. The sensitivity analysis utilized these models to be applied to different meteorological variables to find the best combinations as model inputs to estimate PN concentrations. In this case, sensitivity analysis was found to be more effective than bivariate correlation analysis. Through Pearson correlation coefficient (PCC) analysis, the RH variable was found to have a weak correlation with PN. However, when performing sensitivity analysis, RH was included in the top four best variables for PN modeling. The best combination of measured variables was T, RH, P, and WS. Using these variables as the model inputs, the best modeling techniques were FFNN for daily averaged data with $R^2 = 0.77$ and TDNN for hourly averaged data with $R^2 = 0.66$. The hourly estimation also followed the patterns of the diurnal cycle well, indicating that the established PN model was promising with a satisfactory accuracy.

Nevertheless, this method would be less effective and efficient once the number of measured variables and the amount of datasets become massive. For example, this might take place when the datasets are comprised of more than ten variables or the measurement data resolution is in the scale of minutes for a year-long dataset. Consequently, the method's implementation would be computationally demanding. One solution is to use extra computational resources to parallelize the algorithms, such as through deployment on a computer cluster. Alternative solutions are to implement automatic input selection as proposed in [23] or to apply methods that are capable of performing variable selection, such as LASSO (least absolute shrinkage and selection operator) [42], which have also been used in air pollutant monitoring [43]. Furthermore, since the developed methods were based on statistical methods that utilized data from a specific region, they might work very well in

the training location or in areas with similar emissions, processes, and meteorological conditions. The developed models could be generalized by training them using data measured from different areas. The transfer learning mechanism could be applied to re-train the pre-trained models once there new measurement data are received from the same or a different area [44].

Future works include the use of more time-series models to accommodate time-dependent variables in the form of white-box models, such as dynamic models or black-box models, such as long short-term memory (LSTM). Additional experiments will also be done to include the measurements of trace gases and radiation variables, which might also impact and improve the estimation of PN concentration. Finally, in order to establish a spatio-temporal PN map, more measurements at different locations are needed, and the missing measurements can be done through models' interpolation.

Author Contributions: Conceptualization, M.A.Z., O.S., and T.H.; methodology, M.A.Z. and O.S.; software, M.A.Z. and O.S.; validation, M.A.Z., O.S., and P.L.F.; formal analysis, M.A.Z., O.S., and T.H.; investigation, M.A.Z., O.S., P.L.F., and T.H.; resources, T.H.; data curation, T.H.; writing, original draft preparation, M.A.Z.; writing, review and editing, M.A.Z., O.S., P.L.F., and T.H.; visualization, P.L.F.; supervision, T.H.; project administration, T.H.; funding acquisition, T.H. All authors read and agreed to the published version of the manuscript.

Funding: This research was funded by the Scientific Research Support Fund (SRF, Project Number BAS-1-2-2015) at the Jordanian Ministry of Higher Education and the Deanship of Academic Research (DAR, Project Number 1516) at the University of Jordan. This research was part of a close collaboration between the University of Jordan and the Institute for Atmospheric and Earth System Research (INAR/Physics, University of Helsinki) via ERC advanced Grant No. 742206, the European Union's Horizon 2020 research and innovation program under Grant Agreement No. 654109, the Academy of Finland Center of Excellence (Project No. 272041), ERA-PLANET (www.era-planet.eu), trans-national project SMURBS (www.smurbs.eu, Grant Agreement No. 689443) funded under the EU Horizon 2020 Framework Programme, and Academy of Finland via the Center of Excellence in Atmospheric sciences and NanoBioMass (Project Number 1307537).

Acknowledgments: A part of this research was completed during the sabbatical leave of the last author (Tareq Hussein), which was spent at the University of Helsinki and supported by the University of Jordan during 2019. Open access funding provided by University of Helsinki.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ANN	Artificial neural network
CMAQ	Community multiscale air quality
CO	Carbon monoxide
CPC	Condensation particle counter
FFNN	Feed-forward neural network
LASSO	Least absolute shrinkage and selection operator
LSTM	Long short-term memory
MAE	Mean absolute error
MENA	Middle East and North Africa
NNs	Neural networks
NO_2	Nitrogen dioxide
O_3	Ozone
OPS	Optical particle sizer
P	Absolute pressure
PCC	Pearson correlation coefficients
PM	Particulate matter
PM_{10}	Particulate matter smaller than 10 μm
$PM_{2.5}$	Particulate matter smaller than 2.5 μm
PN	Particle number
R^2	Coefficient of determination
ReLU	Rectified linear unit

RF	Precipitation
RH	Relative humidity
RMSE	Root mean squared error
RNN	Recurrent neural network
SMPS	Scanning Mobility Particle Sizer
SO_2	Sulfur dioxide
T	Temperature
TDNN	Time-delay neural network
UAM	Urban airshed model
UFPs	Ultra-fine particles
WD	Wind direction
WHO	World Health Organization
WS	Wind speed

References

1. WHO Global Ambient Air Quality Database. Available online: https://www.who.int/airpollution/data/en/ (accessed on 19 March 2020).
2. Ayala, A.; Brauer, M.; Mauderly, J.L.; Samet, J.M. Air pollutants and sources associated with health effects. *Air Qual. Atmos. Health* **2012**, *5*, 151–167. [CrossRef]
3. Mannucci, P.M.; Franchini, M. Health effects of ambient air pollution in developing countries. *Int. J. Environ. Res. Public Health* **2017**, *14*, 1048. [CrossRef] [PubMed]
4. Xing, Y.F.; Xu, Y.H.; Shi, M.H.; Lian, Y.X. The impact of PM2.5 on the human respiratory system. *J. Thorac. Dis.* **2016**, *8*, E69. [PubMed]
5. Fuzzi, S.; Baltensperger, U.; Carslaw, K.; Decesari, S.; Denier van der Gon, H.; Facchini, M.C.; Fowler, D.; Koren, I.; Langford, B.; Lohmann, U.; et al. Particulate matter, air quality and climate: Lessons learned and future needs. *Atmos. Chem. Phys.* **2015**, *15*, 8217–8299. [CrossRef]
6. Spinazzè, A.; Fanti, G.; Borghi, F.; Del Buono, L.; Campagnolo, D.; Rovelli, S.; Cattaneo, A.; Cavallo, D.M. Field comparison of instruments for exposure assessment of airborne ultra-fine particles and particulate matter. *Atmos. Environ.* **2017**, *154*, 274–284. [CrossRef]
7. Evans, K.A.; Halterman, J.S.; Hopke, P.K.; Fagnano, M.; Rich, D.Q. Increased ultra-fine particles and carbon monoxide concentrations are associated with asthma exacerbation among urban children. *Environ. Res.* **2014**, *129*, 11–19. [CrossRef]
8. Reche, C.; Querol, X.; Alastuey, A.; Viana, M.; Pey, J.; Moreno, T.; Rodríguez, S.; González, Y.; Fernández-Camacho, R.; Rosa, J.; et al. New considerations for PM, Black Carbon and particle number concentration for air quality monitoring across different European cities. *Atmos. Chem. Physics* **2011**, *11*, 6207–6227. [CrossRef]
9. Frampton, M.W.; Rich, D.Q. Does particle size matter? Ultrafine particles and hospital visits in eastern Europe. *Am. J. Respir. Crit. Care Med.* **2016**, *194*, 1180–1182. [CrossRef]
10. de Jesus, A.L.; Rahman, M.M.; Mazaheri, M.; Thompson, H.; Knibbs, L.D.; Jeong, C.; Evans, G.; Nei, W.; Ding, A.; Qiao, L.; et al. Ultrafine particles and PM2. 5 in the air of cities around the world: Are they representative of each other? *Environ. Int.* **2019**, *129*, 118–135. [CrossRef]
11. Zaidan, M.A.; Wraith, D.; Boor, B.E.; Hussein, T. Bayesian Proxy Modeling for Estimating Black Carbon Concentrations using White-Box and Black-Box Models. *Appl. Sci.* **2019**, *9*, 4976. [CrossRef]
12. Mishra, D.; Goyal, P.; Upadhyay, A. Artificial intelligence based approach to forecast PM2.5 during haze episodes: A case study of Delhi, India. *Atmos. Environ.* **2015**, *102*, 239–248. [CrossRef]
13. Jiang, P.; Dong, Q.; Li, P. A novel hybrid strategy for PM2.5 concentration analysis and prediction. *J. Environ. Manag.* **2017**, *196*, 443–457. [CrossRef] [PubMed]
14. Chang, M.E.; Cardelino, C. Application of the urban airshed model to forecasting next-day peak ozone concentrations in Atlanta, Georgia. *J. Air Waste Manag. Assoc.* **2000**, *50*, 2010–2024. [CrossRef] [PubMed]
15. Mueller, S.F.; Mallard, J.W. Contributions of natural emissions to ozone and PM2.5 as simulated by the community multiscale air quality (CMAQ) model. *Environ. Sci. Technol.* **2011**, *45*, 4817–4823. [CrossRef]

16. Hanna, S.R.; Lu, Z.; Frey, H.C.; Wheeler, N.; Vukovich, J.; Arunachalam, S.; Fernau, M.; Hansen, D.A. Uncertainties in predicted ozone concentrations due to input uncertainties for the UAM-V photochemical grid model applied to the July 1995 OTAG domain. *Atmos. Environ.* **2001**, *35*, 891–903. [CrossRef]
17. Borrego, C.; Monteiro, A.; Ferreira, J.; Miranda, A.; Costa, A.; Carvalho, A.; Lopes, M. Procedures for estimation of modeling uncertainty in air quality assessment. *Environ. Int.* **2008**, *34*, 613–620. [CrossRef]
18. Cabaneros, S.M.S.; Calautit, J.K.; Hughes, B.R. A review of artificial neural network models for ambient air pollution prediction. *Environ. Model. Softw.* **2019**, *119*, 285–304. [CrossRef]
19. García, M.V.; Aznarte, J.L. Shapley additive explanations for NO_2 forecasting. *Ecol. Inform.* **2020**, *56*, 101039. [CrossRef]
20. Nunnari, G.; Dorling, S.; Schlink, U.; Cawley, G.; Foxall, R.; Chatterton, T. Modeling SO_2 concentration at a point with statistical approaches. *Environ. Model. Softw.* **2004**, *19*, 887–905. [CrossRef]
21. Wang, P.; Liu, Y.; Qin, Z.; Zhang, G. A novel hybrid forecasting model for PM10 and SO_2 daily concentrations. *Sci. Total. Environ.* **2015**, *505*, 1202–1212. [CrossRef]
22. Alghamdi, M.A.; Al-Hunaiti, A.; Arar, S.; Khoder, M.; Abdelmaksoud, A.S.; Al-Jeelani, H.; Lihavainen, H.; Hyvärinen, A.; Shabbaj, I.I.; Almehmadi, F.M.; et al. A predictive model for steady state ozone concentration at an urban-coastal site. *Int. J. Environ. Res. Public Health* **2019**, *16*, 258. [CrossRef] [PubMed]
23. Zaidan, M.A.; Dada, L.; Alghamdi, M.A.; Al-Jeelani, H.; Lihavainen, H.; Hyvärinen, A.; Hussein, T. Mutual information input selector and probabilistic machine learning utilisation for air pollution proxies. *Appl. Sci.* **2019**, *9*, 4475. [CrossRef]
24. Fung, P.L.; Zaidan, M.A.; Sillanpää, S.; Kousa, A.; Niemi, J.V.; Timonen, H.; Kuula, J.; Saukko, E.; Luoma, K.; Petäjä, T.; et al. Input-Adaptive Proxy for Black Carbon as a Virtual Sensor. *Sensors* **2020**, *20*, 182. [CrossRef] [PubMed]
25. Raimondo, G.; Montuori, A.; Moniaci, W.; Pasero, E.; Almkvist, E. A machine learning tool to forecast PM10 level. In Proceedings of the AMS 87th Annual Meeting, San Antonio, TX, USA, 14–18 January 2007.
26. Chen, G.; Li, S.; Knibbs, L.D.; Hamm, N.A.; Cao, W.; Li, T.; Guo, J.; Ren, H.; Abramson, M.J.; Guo, Y. A machine learning method to estimate PM2.5 concentrations across China with remote sensing, meteorological and land use information. *Sci. Total Environ.* **2018**, *636*, 52–60. [CrossRef] [PubMed]
27. Mahajan, S.; Chen, L.J.; Tsai, T.C. Short-term PM2. 5 forecasting using exponential smoothing method: A comparative analysis. *Sensors* **2018**, *18*, 3223. [CrossRef]
28. Huang, C.J.; Kuo, P.H. A deep cnn-lstm model for particulate matter (PM2.5) forecasting in smart cities. *Sensors* **2018**, *18*, 2220. [CrossRef]
29. Mølgaard, B.; Hussein, T.; Corander, J.; Hämeri, K. Forecasting size-fractionated particle number concentrations in the urban atmosphere. *Atmos. Environ.* **2012**, *46*, 155–163. [CrossRef]
30. Mølgaard, B.; Birmili, W.; Clifford, S.; Massling, A.; Eleftheriadis, K.; Norman, M.; Vratolis, S.; Wehner, B.; Corander, J.; Hämeri, K.; et al. Evaluation of a statistical forecast model for size-fractionated urban particle number concentrations using data from five European cities. *J. Aerosol Sci.* **2013**, *66*, 96–110. [CrossRef]
31. Hussein, T.; Atashi, N.; Sogacheva, L.; Hakala, S.; Dada, L.; Petäjä, T.; Kulmala, M. Characterization of Urban New Particle Formation in Amman—Jordan. *Atmosphere* **2020**, *11*, 79. [CrossRef]
32. Hussein, T.; Dada, L.; Hakala, S.; Petäjä, T.; Kulmala, M. Urban Aerosol Particle Size Characterization in Eastern Mediterranean Conditions. *Atmosphere* **2019**, *10*, 710. [CrossRef]
33. Pearson, K. *Notes on Regression and Inheritance in the Case of Two Parents*; Proceedings of the Royal Society of London: London, UK, 1895; Volume 58, pp. 240–242.
34. Spearman, C. The Proof and Measurement of Association between Two Things. *Am. J. Psychol.* **1904**, *15*, 72–101. [CrossRef]
35. Zaidan, M.A.; Haapasilta, V.; Relan, R.; Paasonen, P.; Kerminen, V.M.; Junninen, H.; Kulmala, M.; Foster, A.S. Exploring non-linear associations between atmospheric new-particle formation and ambient variables: A mutual information approach. *Atmos. Chem. Phys.* **2018**, *18*, 12699–12714. [CrossRef]
36. Zaidan, M.A.; Canova, F.F.; Laurson, L.; Foster, A.S. Mixture of clustered Bayesian neural networks for modeling friction processes at the nanoscale. *J. Chem. Theory Comput.* **2017**, *13*, 3–8. [CrossRef] [PubMed]
37. Maren, A.J.; Harston, C.T.; Pap, R.M. *Handbook of Neural Computing Applications*; Academic Press: San Diego, CA, USA, 2014.
38. Demuth, H.B.; Beale, M.H.; De Jess, O.; Hagan, M.T. *Neural Network Design*, 2nd ed.; Martin Hagan: Stillwater, OK, USA, 2014.

39. Zaidan, M.; Haapasilta, V.; Relan, R.; Junninen, H.; Aalto, P.; Kulmala, M.; Laurson, L.; Foster, A. Predicting atmospheric particle formation days by Bayesian classification of the time series features. *Tellus B Chem. Phys. Meteorol.* **2018**, *70*, 1–10. [CrossRef]
40. Orhan, U.; Hekim, M.; Ozer, M. EEG signals classification using the K-means clustering and a multilayer perceptron neural network model. *Expert Syst. Appl.* **2011**, *38*, 13475–13481. [CrossRef]
41. Medsker, L.; Jain, L.C. *Recurrent Neural Networks: Design and Applications*; CRC Press: Boca Raton, FL, USA, 1999.
42. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* **1996**, *58*, 267–288. [CrossRef]
43. Van Roode, S.; Ruiz-Aguilar, J.; González-Enrique, J.; Turias, I. An artificial neural network ensemble approach to generate air pollution maps. *Environ. Monit. Assess.* **2019**, *191*, 727. [CrossRef]
44. Pan, S.J.; Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **2009**, *22*, 1345–1359. [CrossRef]